Author comments (AC) **(in bold and blue)** to the referee comments (RC) **(*in bold and italics*)**

General comments:

The study uses interpretable machine learning to identify the most important climate impact drivers for predicting maize and soybean yield variability in Brazilian states. Overall, the manuscript is quite well-written and has clear descriptions of the datasets used which are very helpful for the reader. They discuss in some detail the advantages and disadvantages of the use of different yield datasets in the region, which is crucial for the interpretation of the results of this type of analysis, and make the effort to show a comparison of the datasets and where they agree and disagree. They also use two specific examples of droughts in Brazil as case studies to examine the interpretations, which is interesting and helps to verify their approach.

The topic is very important, and novel methods such as this have a clear use-case in identifying the most relevant periods at which different CIDs impact yields. However, I have some concerns about the methodology. The description of the methodology used is not sufficiently thorough, so these concerns may have been addressed by the authors, but this should be clarified.

AC: We are deeply thankful for the comments of reviewer 1. We will make sure to carefully address each of the suggestions to clarify the concerns about the methodology, so it can be accepted for publication. The suggestion to use spatiotemporal correlation in the cross-validation strategy was the major contribution from reviewer 1 for our work and we will be glad to implement and document this change in our methodology. We are confident that the changes will make our work stronger and increase the impact of the publication in the research community. We provide a more detailed response for each comment in the paragraphs below.

Random forests are often used for this type of study and are a good choice when working with tabular data such as this. However, care must be taken when using any machine learning method not to allow the model to overfit to dependencies or correlation between features. The training and testing method used was not explained clearly, except in Figure 1, which only states that 20% of the data was used to test the models, but not how that 20% was selected. Given that models were trained on a state level, multiple municipalities within each state would have highly correlated climate and yields. Were the data points split in time and/or space to account for this, or sampled randomly? If they were sampled randomly, this can lead to misleading estimations of model performance and the interpretations are less likely to represent the physical mechanisms that are intended to be studied. Particularly relevant - if soil is used as a predictive feature, which does not vary in time in the dataset used (I believe), the model can easily spatially overfit.

Overall, I find the manuscript to be quite well-written and the thorough analysis of the different datasets used and how they impact the results is interesting and excellent scientific practice. However, I think that some small changes to the methodology (most importantly, selecting a test set considering the spatiotemporal autocorrelation and estimating SHAP values using this test set,

ideally using a different feature selection method such as SFS) and better explanation of the steps involved to generate the results discussed could very much improve the paper. As the paper aims to present a framework to enhance the interpretability of ML methods fo crop yield loss prediction, it is important that the framework is robust and can deal with common issues for this type of problem such as overfitting to spatiotemporal data.

AC: We thank referee 1 very much for this comment. Overfitting is a major concern when working with random forests. Each municipality has, at best, ca. 40 data points. This means that there is not enough data for a data-driven model at municipal level. The choice of pooling data at state level is a way of using a priori knowledge to group data. We are pretty aware that this strategy should be further evaluated. However, we also acknowledge that pooling data at state level is an acceptable strategy, especially for policy making. Nearby municipalities can be highly correlated, therefore we understand the concern of referee 1 and the need to further clarify this in the methodology. The papers and methods suggested by referee 1 were very helpful for us to elucidate this issue in our methods.

Finally, given that the title of the paper and stated goal is to present a framework that can be used by other researchers, the code used should be published and made openly available, but this is not currently stated in the manuscript.

AC: We definitely agree with this comment. We are organizing all the scripts to be shared in a repository (e.g., GitHub). This will be appended to the manuscript for the next round of review.

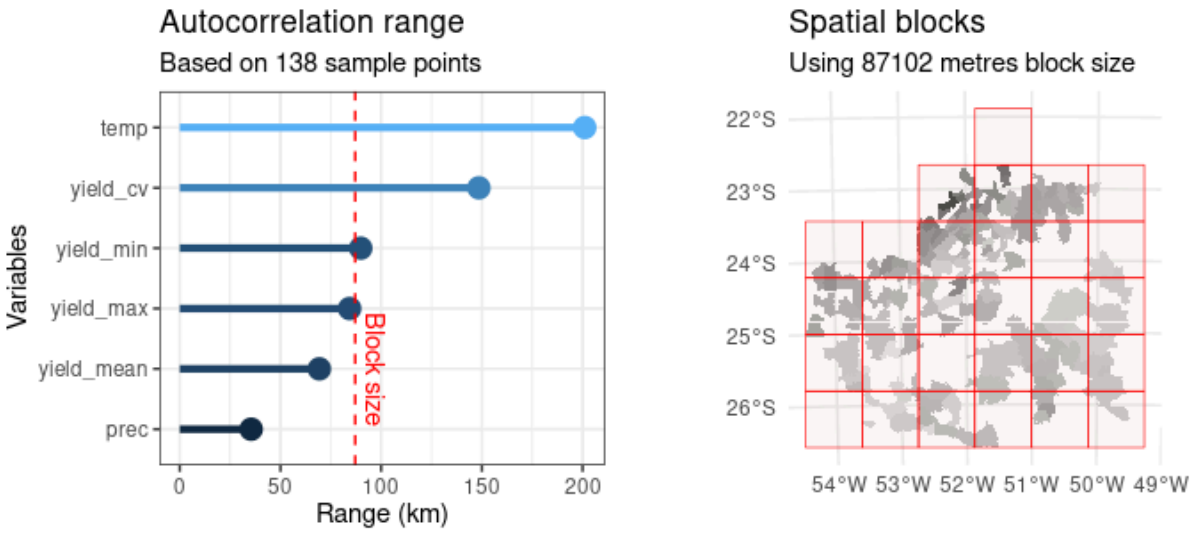
Specific comments:

At what stage was RFE used to select features, and how was this conducted? How many features were selected? I also question the use of RFE in cases where models can overfit (e.g. when spatiotemporal data is used), as features that the model find most important are more likely to not be physically meaningful. Using, for example, sequential feature selection with a spatial or temporal cross-validation splitting method would be more likely to return relevant drivers, and I would recommend to the authors to try this if possible.

AC: Thanks for the suggestions, despite RFE being frequently used for reducing the number of variables, we have not applied this method in our study. We are sorry for that and will make sure to clarify this in the text. In this paper, instead of using RFE, we applied a random forest model to find the 10 most important variables and then we used these variables as input for the SHAP model. After careful consideration of the spatiotemporal overfitting model, we decided to perform important methodological adaptation to clarify this matter and improve the reliability

of our results. These changes can be easily implemented in the R Codes we've already have and should NOT be time costly.

1. **Creating spatial blocks considering spatial autocorrelation of crop yields using Valavi et al. (2019) that applies Roberts et al. (2017) cross-validation strategy. From this, we can create spatial blocks that incorporate autocorrelation of crop yields, temperatures and precipitation means over municipalities. Here, we present an example that considers that autocorrelation has a range of 87 km for the Deral dataset in the Brazilian state of Paraná.**



2. **Spatio-temporal k-fold Cross-validation using the method of nearest neighbor distance matching (Mila et al 2022, Linnenbrink et al 2023) available for implementation in R using the CAST package (Meyer et al., 2024).**
3. **Training a random forest model using the caret package (Kuhn, 2022);**
4. **Removing variables that cause overfitting using forward feature selection. It is important to say that this step will substitute the need of selecting only the 10 most important variables;**
5. **Using the most important variables in the SHAP framework.**

In Figure 4, it would be helpful to have descriptions of what features were included in the different scenarios - in particular, I could not understand what 'Complex' meant.

AC: We thank referee 1 for the comment and we will make sure to add the descriptors suggested in Figure 4. There are multiple ways to derive multi-hazard scenarios for crop yields. As a way to promote a simplification focusing on improving the explainability of the different models, we decided to sub-divide input climate drivers according to the hazard type, i.e., precipitation means only, temperature means only, precipitation and temperature means, and

precipitation means, temperature means and extreme indices. The combination of mean weather variables and extreme weather variables was called 'Complex'. We will make sure to make it more clear in the methodology section describing the scenarios using a table.

In Figure 5 and 6, is this after RFE has been used to select only 10 features? I was confused by the fact that for maize, only February features are shown, but later in the text it states that April and May precipitation was important for some regions.

Thank you for your comment. In fact, the maize second cycle growing season starts in February, therefore, the beginning of the growing season plays an important role in determining the crop yields. We should make it clear in the discussion of the results.

I would strongly advice not removing correlated variables before doing the feature selection. You can expect that the highly correlated variables will not both be selected, and it is another opportunity for data leakage to enter.

We appreciate the comment and agree that we can avoid removing correlated variables, since we perform random feature elimination, we should use all the variables in our study avoiding data leakage.

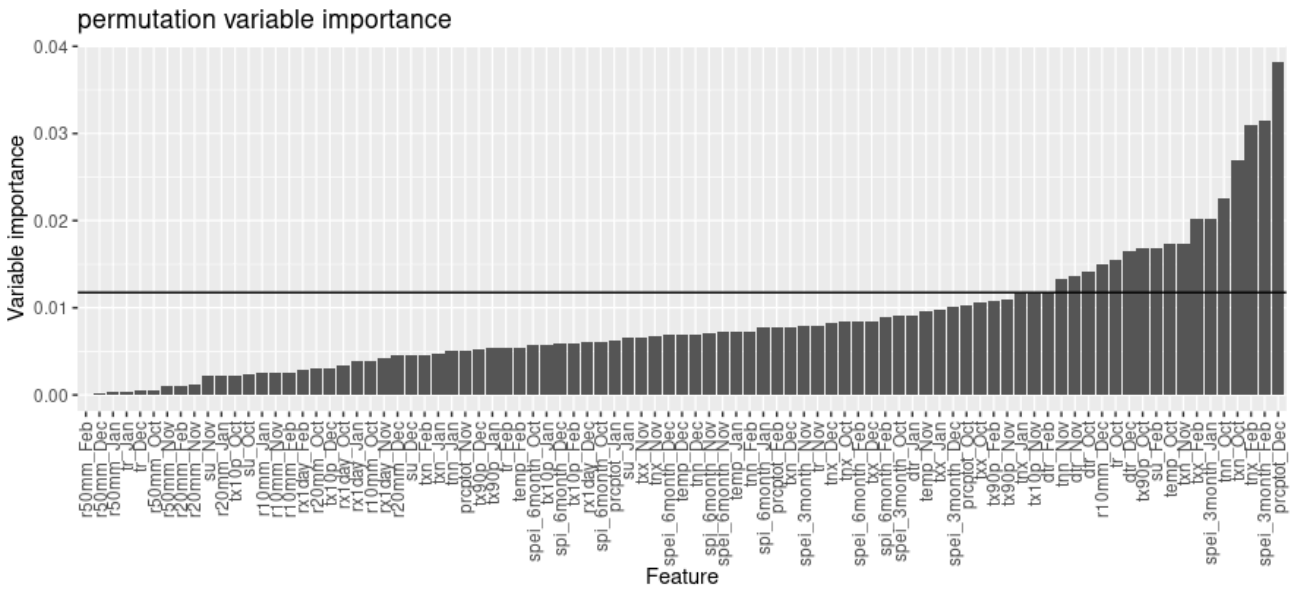
I think it is very useful to compare the importances between the different states and datasets, as this can help to find robust insights and identify potential problems with the datasets used. It would be useful to see uncertainty quantification here as well, as given that similar model performance can come from many combinations of features (as shown in Figure 4), one would expect that there is significant uncertainty in the interpretations as well. I would also consider using an additional feature importance metric (permutation feature importance on held-out test set?) for comparison, but this might be out of scope.

We agree with the comment and think that the held-out test set could help us to improve the discussion with uncertainty quantification. While this suggestion can provide additional feature importance metrics, we believe that this could be explored in future work.

I also find it unusual to fit random forest models and then to use a more complex model (XGBoost) to explain them via SHAP. Normally, SHAP is used directly on the trained model to be interpreted, and if a second model was used it would normally be a simpler model. Why not use XGBoost for the initial part of the analysis instead of adding this complexity of using a second model to explain the first?

Thank you for your comment. We agree that this choice should be better explained in the manustrip. SHAP uses game theory to analyze the impact of variables considering the interaction with other variables. In other words, Shapley values estimate the importance of a

feature by contrasting the model prediction without that feature. The computation cost of adding all the variables makes running the model in a modern microcomputer impractical. Moreover, since we are decomposing the indices in monthly steps, in order to make the *post hoc* analysis feasible, we need to reduce the dimensionality of input variables. We understand that choosing 10 variables as a prior definition can exclude important variables. Since we will perform a permutation feature importance, we can use a posteriori criteria such as the input variables that represent ca. 80% of model variability.



Partial dependence plots do not need SHAP values - they can be calculated by just varying individual features and estimating the output. It might be interesting to compare this against those gained from SHAP (but again, maybe out of scope). It would at least be useful to discuss/justify in the text why the partial dependence plots gained from SHAP are more useful (which is very plausible).

We thank the reviewer for the comment. Partial dependence plots were used in other papers (e.g., Vogel et al., 2019) that were cited in our manuscript. We agree that it is an interesting idea to compare the two plots to understand what insights can be gained by the use of SHAP partial dependence plots, however, we believe that this comparison should be included in future work.

SHAP values are also sensitive to the data used to calculate them, and I would again recommend to use test sets for this that are split with consideration to the spatial and temporal correlations.

We thank you for the comment. As described previously, the spatial and temporal correlations will be taken into consideration for the improved model and we believe that this will significantly improve the reliability of our results.

Interpreting the results of this type of study can be difficult, as in general, any feature used for training is one that could be a causal driver. This means that it is hard to figure out if the results are meaningful or if the model has learned some spurious correlations. The fact that only February features are shown as important for maize suggests, to me, that something strange is going on, as the authors state that this is peak planting date and in some regions, planting is not finished until the beginning of April. It seems more likely that heat, for example, would be more important during the reproductive period. Using the different test sets as I mentioned before might help with this, as well as using instead of the internal RF variable importance measure.

We agree with the reviewer that interpreting the results is difficult, especially considering the complex systems like climate and agriculture. We need to review these results. To enhance the robustness and interpretability of our results, we will apply the permutation feature importance test in addition to the current methodology. This adjustment will help validate our findings and provide a clearer understanding of the causal relationships driving maize yields, particularly in the context of varying climate conditions.

Why remove heteroskedasticity? Could this be justified more in the text? As we expect more climate variability with climate change and therefore more yield variability, it isn't obvious that this should be corrected for.

We thank the reviewer for this comment and we believe that the influence of climate change on crop yield variability is a complex topic. We based the decision to remove heteroskedasticity on the actuarial literature (Tolhurst and Ker, 2014; Liu and Ker, 2020; Osaki et al., 2008). The aforementioned papers demonstrate the importance of removing heteroskedasticity and that it can be also related to technological changes.

Lines 171-172 describe a second analysis using Gaussian copulas, but I could not find this further described or any results from this in the rest of the manuscript?

In the manuscript, we mention a second analysis that employs Gaussian copulas to evaluate the combined effect of variables. This analysis complements the analysis using SHAP values to interpret the contributions of individual features. SHAP values provide a way to quantify the contribution of each feature to the prediction made by a model. The combination of SHAP values and Gaussian copulas, allowed us to further evaluate the dependencies between features. In other words, we evaluated the effect of one feature might be influenced by the presence or value of another feature. We agree with the reviewer that this segment should be better explained both in the methodology section and in the results section.

Technical corrections:

The paragraph on interpretability (lines 53 to 56) I could not understand.

We thank the reviewer for the comment. We agree that the paragraph needs to be improved both in terms of logic and language.

Old version:

The paradigm of interpretability of machine learning models is a broad topic of discussion in supervised learning (Lipton, 2018). Two essential observations related to model interpretability are: (i) and (ii) the training data can be imperfect to represent a dynamic environment that changes over time.

New version:

The paradigm of interpretability of machine learning models is a broad topic of discussion in supervised learning (Lipton, 2018). The model interpretability can be achieved by means of feature engineering and using interpretable models such as linear models, that is "algorithmic transparency". When the features, or input data, are decomposed and the number of variables make the interpretation of models difficult, post hoc interpretation can be used to extract explanations from learned models.

Please state briefly that the crop yields were detrended in the main text (the further explanation in the Supplementary is very helpful, but there is no mention of the fact that the yields are detrended in the main manuscript which is very important to interpret the results).

We agree with the comment. We can bring the main information that was handed in the supplementary material to the main text to better explain

Some references on selecting test sets appropriately when using ML with spatiotemporal data:

Meyer, H., Reudenbach, C., Wöllauer, S. & Nauss, T. Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. Ecological Modelling 411, 108815 (2019).

Sweet, L., Müller, C., Anand, M. & Zscheischler, J. Cross-Validation Strategy Impacts the Performance and Interpretation of Machine Learning Models. Artificial Intelligence for the Earth Systems 2, (2023).

Roberts, D. R. et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929 (2017).

We thank for the references suggested by the reviewer and we believe that adopting the spatial and temporal aspects mentioned in the text will considerably improve the impact of the manuscript.

References

Kuhn. caret: Classification and Regression Training. R package version 6.0-93. 2022. Available at https://CRAN.R-project.org/package=caret.

Meyer et al.. CAST: 'caret' Applications for Spatial-Temporal Models. R package version 1.0.1 2024. Available at https://CRAN.R-project.org/package=CAST.

Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G. blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods Ecol Evol. 2019; 10:225–232. doi: 10.1111/2041-210X.13107.